Cataloguer Support in Cultural Heritage

Gazendam, L.J.B.

2015

document version
Publisher's PDF, also known as Version of record

Samenvatting Ondersteuning van Documentalisten binnen het Cultureel Erfgoed

Bibliotheken, archieven en musea verzamelen schilderijen, beelden, boeken, muziek, films en andere objecten. Documentalisten beschrijven[18] eigenschappen van de objecten, zoals de titel, de makers, het onderwerp etc. Deze beschrijvingen dienen twee doelen: 1) de verzameling toegankelijk maken (gebruikers kunnen de collecties doorzoeken) en 2) objecten duiden: een object een bepaalde betekenis geven (door het bijvoorbeeld te classificeren als *communistische kunst*).

Het Nederland Instituut voor Beeld & Geluid is het Nederlandse nationale audiovisuele archief. Door digitalisering is de omgeving van Beeld & Geluid radicaal veranderd. De verminderde productie- en opslagkosten van audiovisueel materiaal resulteren in een enorme toename van instromend en opgeslagen materiaal (in 2014 werd ongeveer 40.000 uur nieuw materiaal opgeslagen). Naast de toename van het instromende materiaal, verandert digitalisering het speelveld voor Beeld & Geluid. Sinds de Nederlandse publieke omroepen zijn overgestapt op een volledig digitale infrastructuur, verwachten catalogusgebruikers, zoals professionele omroepen, dat nieuw materiaal op de dag van de uitzending (gedocumenteerd) in de online catalogus verschijnt.

Handmatige annotatie is een tijdrovende bezigheid voor de 42 documentalisten die bij Beeld & Geluid werkzaam zijn. Ze slagen erin om ongeveer eenderde van al het jaarlijks instromende materiaal te beschrijven, waarbij de definitieve omschrijving ongeveer twee maanden na de uitzending klaar is. De beschrijving van de overige tweederde bestaat slechts uit de gegevens waarmee het werd aangeleverd aan het archief, meestal is dat de titel plus technische informatie zoals uitzenddatum, -tijd en het -kanaal.

Bij Beeld & Geluid gebruiken documentalisten twee hulpmiddelen bij het indexering: iMMix en GTAA. iMMix is Beeld & Geluids archiveringssysteem. iMMix is opgesplitst in twee delen: het iMMix metadatasysteem slaat de metadata op en Het Digitale Archief slaat de essense op (het echte beeld- en geluidsmateriaal). De iMMix metadata bevat o.a. technische informatie (zoals lengte, uitzenddatum), inhoudsbeschrijvingen (zoals trefwoorden, locaties, personen, makers, producenten) en informatie over digitale rechten. De meeste inhoudsbeschrijvers, zoals trefwoorden, locaties,

[18] Ook wel annoteren, indexeren of metadateren genoemd.

personen, makers en producenten, zijn beperkt tot Beeld & Geluids eigen thesaurus GTAA (GTAA staat voor Gemeenschappelijke Thesaurus Audiovisuele Archieven). De GTAA bevat ongeveer 160.000 termen, georganiseerd in 6 disjuncte facetten: Onderwerpen, Genres, Personen, Namen[19], Makers en Plaatsen.

Termen uit alle facetten van de GTAA kunnen relaties bevatten, zoals *Gerelateerde Term*[20] en *Afbakening Notities*[21], alleen Trefwoorden en Genres kunnen ook *Gebruik / Gebruik voor*[22] en *Bredere Term / Smallere Term*[23] relaties hebben. Met deze laatste twee kunnen termen in hiërarchieën georganiseerd worden.

Dit maakt onze onderzoeksvraag: **Hoe kunnen cultureel erfgoed documentalisten automatisch worden ondersteund tijdens de creatie en de verrijking van catalogus annotaties?**

Het is onze hypothese dat de gestructureerde achtergrondinformatie, zoals opgeslagen in thesauri als de GTAA en in de catalogusbeschrijvingen, nuttig is voor het ondersteunen van de documentalisten gedurende de taak van het indexeren. We hanteren een ingenieursbenadering en ontwerpen (automatische) processen en bouwen prototypes. Deze worden vervolgens geëvalueerd tijdens experimenten.

In Hoofdstuk 2: *Webgebaseerde Thesaurus Browsing* beschreven we de ontwikkeling en evaluatie van een thesaurusbrowser. In de thesaurusbrowser hebben we getracht de verschillende aspecten van de GTAA toegankelijk te maken: de facetten, de termen, hun relaties, hun hiërarchieën en de facetcategorieën. We hebben ons ontwerp getest en aangepast tijdens twee rondes van gebruikersevaluaties. Tijdens deze experimenten voerden documentalisten en archiefgebruikers een indexeringstaak uit met behulp van de thesaurusbrowser. De Beeld & Geluid documentalisten, die gewend waren relatief basale instrumenten te hanteren om termen te vinden, werden overdonderd door de complexiteit van de browserinterface. Ze waren gewend alfabetisch te zoeken en daarom hadden we het meeste baat bij het optimaliseren van de alfabetische zoekfunctie in de browser. Om termen te vinden werd niet veel gebruik gemaakt van het navigeren in hiërarchische relaties. Voor het bepalen van de precieze betekenis was de hiërarchie juist wel handig: door de semantische omgeving van de term te bekijken, kon bepaald worden of een term de juiste was[24]. We hebben gemerkt dat daar ook het grootste verschil bestond tussen de documentalisten van Beeld & Geluid (deskundigen op het gebied van de thesaurusinhoud) en de gebruikers van de omroepen. De gebruikers van de omroepen waren veel meer geneigd om te zoeken naar een term door met bladeren in hiërarchieën te beginnen. Deze browse-optie kan nuttig zijn om de dagelijkse gebruikers van de Beeld & Geluid archieven verder te ondersteunen.

De documentalisten waren over het algemeen positief over het gebruik van het ontworpen instrument voor hun dagelijks werk. Dit blijkt niet alleen uit de vragenlijst, maar ook uit het feit dat mede op basis van de resultaten van deze studie Beeld & Geluid aspecten van de thesaurusbrowser in hun archiveringsproces heeft ingebouwd.

[19] Voor namen van organisaties, bewegingen en andere namen.
[20] Een *Gerelateerde Term* legt bijvoorbeeld een relatie tussen *kernenergie* en *kernreactor*.
[21] Uitleg over hoe de term te gebruiken.
[22] De *Gebruik*-relatie verwijst bijvoorbeeld *atoomenergie* door naar de voorkeursterm *kernenergie*.
[23] *Atoomenergie* is een *Smallere Term* van *energie*.
[24] De term *missie* is gerelateerd aan *katholieke kerk* en *ontwikkelingssamenwerking*. Door deze te tonen in de browser kon een gebruiker snel beslissen of een term juist is.

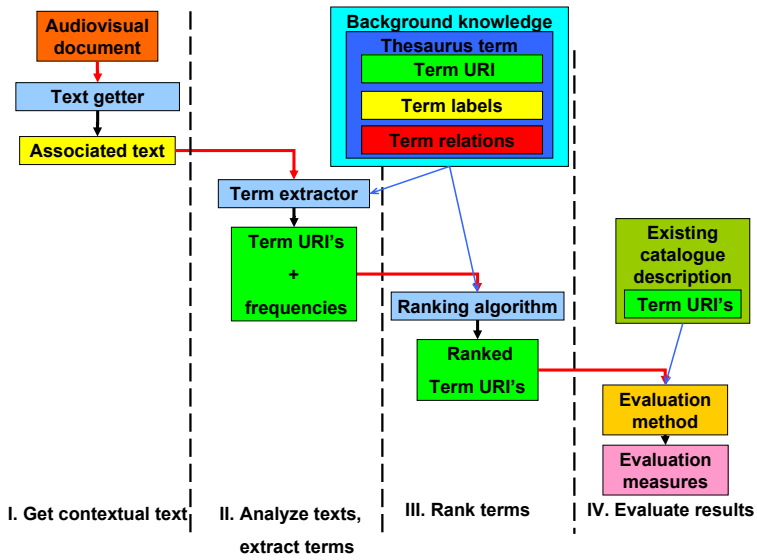

Figure A.2: De vier fasen bij het automatisch genereren van geordende thesaurustermen voor audiovisuele documenten.

In Hoofdstuk 3: *Automatisch Annotaties Genereren* beschreven we een prototype dat we ontworpen hebben om automatische annotatiesuggesties te genereren. Het prototype kreeg als invoer de GTAA en sets teksten die gerelateerd waren aan de te archiveren objecten. De architectuur voor dit systeem wordt geschetst in Figuur A.2.

Het prototype analyseerde contextuele informatie, in dit geval waren dat vier websiteteksten die behoorden bij twee TV-programma's. Uit deze teksten werden (GTAA) termen geëxtraheerd en met behulp van de structuur in de GTAA geordend (belangrijkste termen bovenaan). De uitkomsten van het prototype waren zes soorten gerangschikte annotaties. We evalueerden de uitkomsten door deze te vergelijken met a) de beschrijvingen van de twee programma's gemaakt door Beeld & Geluid experts (managers kwaliteitsbewaking) en b) de beschrijvingen die negen verschillende documentalisten gemaakt hadden van de twee programma's in een gebruikersstudie. De vergelijking met de beschrijvingen van de negen documentalisten gaf inzicht in de consensus tussen de documentalisten. Uit beide evaluaties kwam naar voren dat sommige annotaties wel goed waren, maar niet exact overeenkwamen met de menselijke annotaties (bijvoorbeeld wanneer *De Kaukasus* werd voorgesteld en *Tsjetsjenië* en *Dagestan* voorkwamen in de menselijke annotaties[25]). Om dit probleem op te lossen, introduceerden we semantische evaluatie in Hoofdstuk 4.

[25]De Kaukasus is de bergketen tussen Tsjetsjenië en Dagestan waar de makers hun programma filmden.

In Hoofdstuk 4: *Evaluatie van Automatische Annotaties* hebben we het protoype uit Hoofdstuk 3 volledige geïmplementeerd en richtten we ons op de evaluatie van de uitkomsten[26].

We hebben een corpus van 258 uitzendingen gecreëerd waarvoor we automatische annotaties afleidden. Op de resultaten hebben we een klassieke informatie-extractie evaluatie toegepast door onze uitkomsten te vergelijken met de bestaande catalogus beschrijvingen. Bij nadere beschouwing toonde deze evaluatie echter tekortkomingen. Op twee manieren hebben we de tekortkomingen nader onderzocht: ten eerste door de evaluatie te veranderen naar semantische evaluatie, ten tweede door de invoering van het idee van serendipitous browsen.

De lossere **semantische evaluatie** rekent ook de termen goed die één thesaurus relatie verschilden van het handmatig toegekende catalogus trefwoord. Daarmee is het in staat om kleine semantische verschillen die tot ongewenste fouttelling leiden op te vangen. De automatisch afgeleide annotaties *ministers, varkenspest, landbouw, ministeries* en *varkens* bijvoorbeeld, zijn allemaal fout wanneer deze precies vergeleken worden met de handmatige catalogus trefwoorden *pest, vee, vaccinaties* en *veterinaire ziekten*. Wanneer we deze semantisch evalueren worden *varkenspest, landbouw* en *varkens* wel als correct beoordeeld. In het licht van de indexeringstaak van documentalisten en de conceptuele modellering van de GTAA thesaurus is overstappen naar een semantische evaluatie een goede strategie bij het evalueren van automatische indexeringstechnieken.

Bij het **serendipitous browsen** worden (automatische en handmatige) annotaties gebruikt om potentieel interessante gerelateerde documenten te vinden. Het experiment met serendipitous browsen toonde aan dat de automatische annotaties en de catalogusannotaties dezelfde waarde hebben voor het vinden van interessante gerelateerde documenten.

In Hoofdstuk 5: *Automatische Trefwoordsuggestie* bestudeerden we afzonderlijk verschillende realisaties van en verschillende inputs voor de vier onderdelen van het automatische annotatieproces.

In Sectie 5.2 bestudeerden we de ordeningalgoritmes. We hebben gevonden dat de ordeningsalgoritmes de uitkomsten van de trefwoordsuggestie beïnvloeden. Ons beste algoritme was het TF.RR algoritme, dat zowel de thesaurusrelaties als de frequentie informatie (van de in de teksten gevonden termen) gebruikt. Het was even goed als het klassieke TF.IDF algoritme, maar heeft in tegenstelling tot dit klassiek algoritme geen achtergrondcorpus nodig.

In Sectie 5.3 hebben we verschillende soorten gestructureerde achtergrondinformatie (zoals de GTAA) bestudeerd. De handgemaakte GTAA was duidelijk het nuttigst. Het leidde tot veel betere resultaten dan het co-occurrence netwerk dat afgeleid is van de iMMiX catalogus beschrijvingen. Dit netwerk relateert trefwoorden op basis van hoe vaak ze samen gebruikt worden. De gedachte was dat gemeenschappelijk gebruik een gemeenschappelijke betekenis impliceert.

In Sectie 5.4 varieerden we de contextinformatie (waaruit we de trefwoorden extraheren). De eerder genoemde contextdocumenten werden gebruikt en automatische spraakherkenning (ASR). Een experiment toonde aan dat de ARS en de contextdoc-

[26] We hebben drie verschillende ordeningsalgoritmes geïmplementeerd voor stap III in Figuur A.2.

umenten beide waardevolle en complementaire informatiebronnen zijn. De ASR leek wel iets minder geschikt als bron om annotaties uit af te leiden.

In Sectie 5.5 hebben we de interactie tussen automatische annotatie en zoekvraagverbreding bestudeerd. Het experiment toonde aan dat automatische annotaties gemiddeld iets minder waardevol dan handmatige annotaties, maar dat er bijna geen overlap is tussen beide. Doordat in veel gevallen de zoekvraagverbreding tot een verbetering van het aantal gevonden documenten leidde, zonder veel extra irrelevante documenten op te leveren, zijn beide bronnen waardevol. De structuur van de thesaurus echter, zorgt ervoor dat de zoekvraagverbreding voor sommige termen wel veel irrelevante documenten oplevert[27] en bij andere termen niets uitmaakt[28].

In Hoofdstuk 6: *Documentalist Ondersteuningssysteem* integreerden we automatische annotaties en de visualisatie van de informatieomgeving in een documentalisten ondersteuningssysteem (DocSS). In het DocSS kan een documentalist allerlei informatie bekijken die tijdens het catalogiseren van een afzonderlijk object relevant is. Het toont informatie op collectieniveau, bijvoorbeeld alle documenten bij één serie: de verschillende catalogusbeschrijvingen en alle bijbehorende (website)teksten. Het toont ook informatie bij afzonderlijke afleveringen, alle bij één aflevering behorende contextdocumenten, in elk contextdocument worden alle GTAA termen gehighlight, uit al deze gehighlighte termen wordt automatisch een gerangschikte lijst van annotaties gemaakt. Binnen deze omgeving kan ruim en smal gezocht worden: een gebruiker kan bijvoorbeeld zoeken naar *Arnhem* en vervolgens kiezen of hij alleen resultaten wil krijgen waarbij *Arnhem* in de catalogusbeschrijving voorkwam, of ook binnen de annotatie-suggesties of in de tekst van een contextdocument.

Hoewel dit prototype niet in een archiveringstaak getest is, bood het interessante mogelijkheden voor semantisch zoeken en navigeren in de rijke omgeving van de Beeld & Geluid catalogus. De automatische koppeling aan thesaurustermen en aan vergelijkbare documenten, het creëren van navigatielinks die de informatieomgeving tonen, de hints naar interessante gerelateerde informatie (informatiegeur [136]), het interactieve karakter: alles samen geeft veel ideeën voor nieuwe sterk interactieve gebruikersinterfaces voor informatieinteractie [136].

In het laatste onderzoekshoofdstuk, Hoofdstuk 7: *Gebruikers Evaluatie van Handmatige en Automatische Annotaties*, bestudeerden we in een gebruikersstudie de verschillen en de overlap tussen handmatige en automatische annotaties. In deze evaluatie hebben we verschillende gebruikersgroepen laten oordelen hoe goed annotaties zijn. Dit oordeel ging over zowel handmatige als automatische annotaties. De drie groepen waren documentalisten, beroepsmatige archiefgebruikers en geïnteresseerde buitenstaanders.

Bij het bekijken van de overlap en de verschillen tussen de handmatige en automatische annotaties in paragraaf 7.4, zagen we dat 40% van de als goed beoordeelde annotaties zowel door de documentalisten als door de automatische technieken gegenereerd wordt. Nog eens 40% vonden we alleen bij de automatische technieken en 20% werd uitsluitend gegenereerd door de documentalisten. De goede automatische annotaties werden echter vergezeld door een zeer groot aantal slechte annotaties, terwijl de hand-

[27] Bij de termen met zeer veel relaties leidde de zoekvraagverbreding snel tot slechte resultaten.
[28] Bij termen zonder relaties helpt het niets.

matige annotaties bijna allemaal goed waren.

Dit wil niet zeggen dat deze resultaten nutteloos zijn voor een documentalist. Van de automatische annotaties waren de resultaten het slechtst voor het type *makers* (slechts eenderde van de automatische suggesties was daar correct). Als we deze automatische annotaties aanbieden aan een documentalist, verandert de taak van de documentalist van het kiezen van 6 *makers* uit een lijst van 18.000 in het kiezen van 5 *makers* uit een set van 20. Dit kan het aantal annotaties en de werksnelheid van een documentalist verhogen. Een onderzoek naar het effect van automatische voorbeschrijvingen op de werksnelheid van een documentalist toont een verbetering in de werksnelheid van 25-35% [46] aan.